Cell Research (2010) 20:508-509.

www.nature.com/cr

RESEARCH HIGHLIGHT

Decoding the dual-coding region: key factors influencing the translational potential of a two-ORF-containing transcript

Han Liang[1]

[1]*Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. Houston, TX 77030, USA*

Cell Research (2010) **20**:508-509. doi:10.1038/cr.2010.62; published online 3 May 2010

A dual-coding region is defined as a stretch of DNA that encodes amino acids in overlapping reading frames. Dual-coding regions are often found in bacteriophages and viruses (e.g., HIV) with tiny genome sizes; such an arrangement is believed to greatly increase genetic information storage efficiency [1]. In mammals, genetic information storage is not an issue because the mammalian genome is huge and contains large amounts of non-coding sequences. Coding regions usually encode amino acids only in one reading frame, but there are some exceptions. Generally speaking, dual-coding regions can arise from three sources: (1) from nearby overlapping genes (two genes can be either in the same strand or in opposite strands); (2) from alternatively spliced transcripts of the same gene; and (3) from different translational initiation sites of a single transcript (Figure 1). Dual-coding regions have recently attracted wide interest [2-4]. The identification and characterization of these special coding regions will improve the current gene/genome annotation, and contribute to a deeper understanding of the protein translation mechanism.

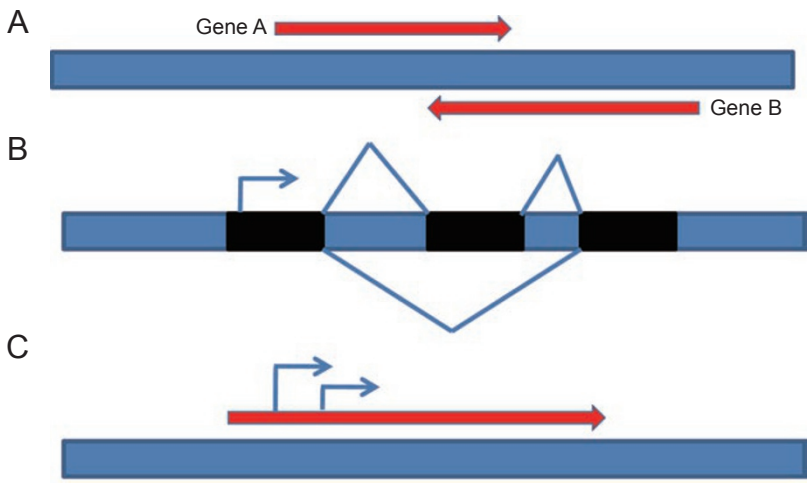

Figure 1 Three types of dual-coding regions: **(A)** from nearby overlapping genes (two genes can be either in the same strand or in opposite strands); **(B)** from alternatively spliced transcripts of the same gene, where the black boxes indicate exons; and **(C)** from different translational initiation sites of a single transcript.

Most intriguing among the three sources of dual-coding regions mentioned previously is that which arises from different translational initiation sites on a single transcript. This construct involves the generation of two distinct protein products from the same mRNA transcript. Although current bioinformatic approaches can readily identify hidden reading frames in annotated frames and can define transcripts that contain two open reading frames (ORFs) [2], without experimental evidence at the translational level, the amount and significance of true dual-coding regions will remain unclear. A fundamental question is what factors determine the on/off function of the two peptide products potentially encoded in a two-ORF-containing transcript. Xu and colleagues address this question in a recently published article [5], in which they describe an elegant *in vitro* dual-coding expression system they developed. In their study, two fluorescent proteins (green fluorescent protein [GFP] and red [RFP]) were encoded in an artificial transcript. The

Correspondence: Han Liang
Tel: 1-713-745-9815; Fax: 1-713-563-4242
E-mail: hliang1@mdanderson.org

GFP (longer ORF) and RFP (shorter ORF) started with the first AUG codon (the start codon) in two overlapping ORF configurations, respectively. For each configuration, the combinations of strong and/or weak Kozak motifs [6] were respectively introduced to flank the two AUG codons. Thus, Xu *et al*. [5] were able to systematically examine the impact of three key factors on the translational activity of a two-ORF-containing transcript: the position of the start codon, the ORF length and the strength of the Kozak motif. Using fluorescence imaging techniques and western blots, they detected dual protein products in four out of the eight transcript subtypes under survey. Among the ORFs, the vast majority (7/8) of ORFs with the first AUG and the longer ORFs were translated. For the first time, these results have established a set of explicit rules for predicting whether a two-ORF-containing transcript can generate two protein products simultaneously.

Using the rules inferred from their experimental studies, Xu *et al*. [5] further performed a bioinformatic analysis to screen the dual-coding transcripts in the human and mouse genomes. They found that about 170 human transcripts are potentially dual coding and only 18 are conserved in the mouse genome. When the effect of nonsense-mediated mRNA decay was considered, those numbers were reduced to 80 and 9, respectively. Only a relatively small percentage of dual-coding transcripts are conserved between the two species, suggesting a recent origin of most dual-coding transcripts. This observation also implies that the evolutionary processes that involve these special transcripts are quite dynamic, which is similar to our understanding of microRNAs [7]. Conducting additional translation on an already-existing transcript may provide an energetically-efficient way to generate a pool of raw materials within the evolutionary process. But such an overlapping reading fame arrangement comes with a cost: the constraints of an existing reading frame do not allow for free exploration of the amino acid space in a second reading frame. Thus, a new protein product is only occasionally selectively favored and maintained in long-term evolution.

The study of Xu *et al*. [5] provides crucial insights into characterizing dual-coding transcripts in mammalian genomes, and raises some interesting questions. First, to what extent can the observation based on the artificial dual-coding transcript be generalized to other transcripts? In the experiment, the first and second AUG codons are only a few nucleotides away, and the GFP ORF is about two times longer than the RFP ORF. These parameters often vary greatly from transcript to transcript, and so the real predictive power of the proposed model remains to be seen. Moreover, the rules obtained from the study are in a qualitative format, but the relative expression levels of the two protein products show significant variations among different transcript subtypes. Hence, it would be desirable to build a quantitative model to more accurately predict the translational behavior of a dual-coding transcript. Finally, the *in vitro* study, in a sense, only examines the translational potential of a transcript in a simplified environment. The translation of cellular messenger RNA *in vivo* is a more complex process and often involves many factors that are specific to a given tissue or cell type. Nevertheless, Xu *et al*. [5] obtain a set of experimentally-determined and biologically-sensible rules to identify dual-coding transcripts, which provides a valuable starting point for further investigation.

References

1 Normark S, Bergstrom S, Edlund T, *et al.* Overlapping genes. *Annu Rev Genet* 1983; **17**:499-525.
2 Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol* 2007; **3**:e91.
3 Liang H, Landweber LF. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res* 2006; **16**:190-196.
4 Tress ML, Martelli PL, Frankish A, *et al.* The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci USA* 2007; **104**:5495-5500.
5 Xu H, Wang P, Fu Y, *et al.* Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res* 2010; **20**:445-457.
6 Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol Cell Biol* 1989; **9**:5073-5080.
7 Liang H, Li WH. Lowly expressed human microRNA genes evolve rapidly. *Mol Biol Evol* 2009; **26**:1195-1198.